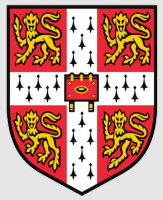


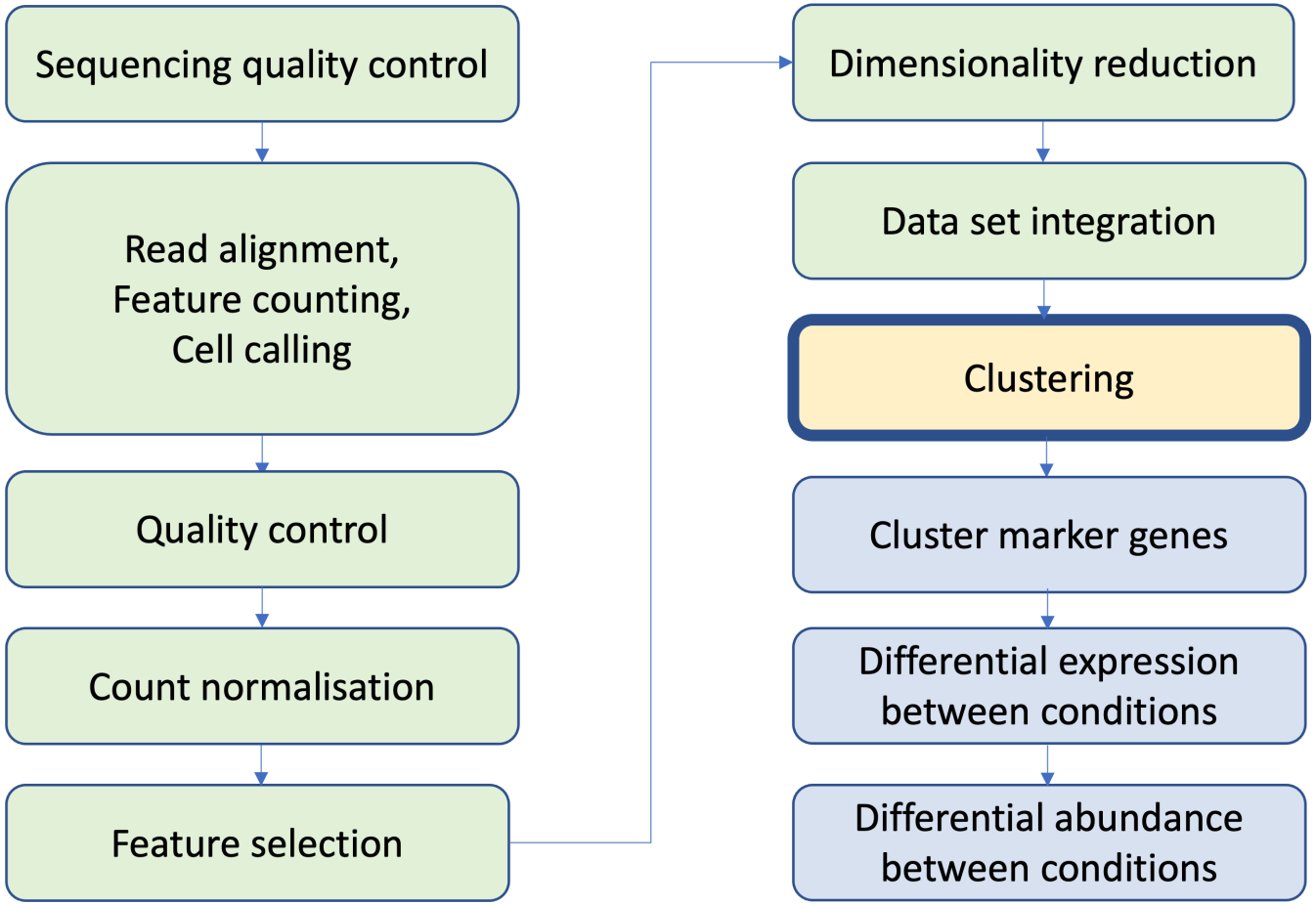
UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK | Cambridge Institute

Clustering

February 2026

Single Cell RNAseq Analysis Workflow



Motivation

The data has been QC'd, normalized, and batch corrected.

We can now start to understand the dataset by identifying cell types. This involves two steps:

- unsupervised clustering: identification of groups of cells based on the similarities of the transcriptomes without any prior knowledge of the labels usually using the PCA output
- annotation of cell-types based on transcription profiles

Clustering methods

- Roughly classified into four categories
  - k-means clustering
  - hierarchical clustering
  - density-based clustering
  - Graph-based clustering
- First three methods dose not scale well for the large data sets.
- Data from single cells is best clustered using a graph-based approach, as it is faster and more efficient.

Graph-based clustering

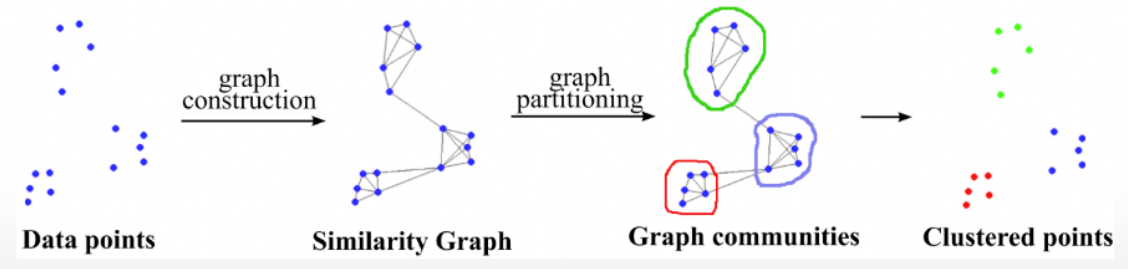
Pros

- fast and memory efficient (no distance matrix for all pairs of cells) compared to hierachical clustering
- no assumptions on the shape of the clusters or the distribution of cells within each cluster compared to e.g. k-means or gaussian mixture models
- no need to specify a number of clusters to identify

Cons

- loss of information beyond neighboring cells, which can affect community detection in regions with many cells.

The steps involved:



Making a graph

Nearest-Neighbour (NN) graph:

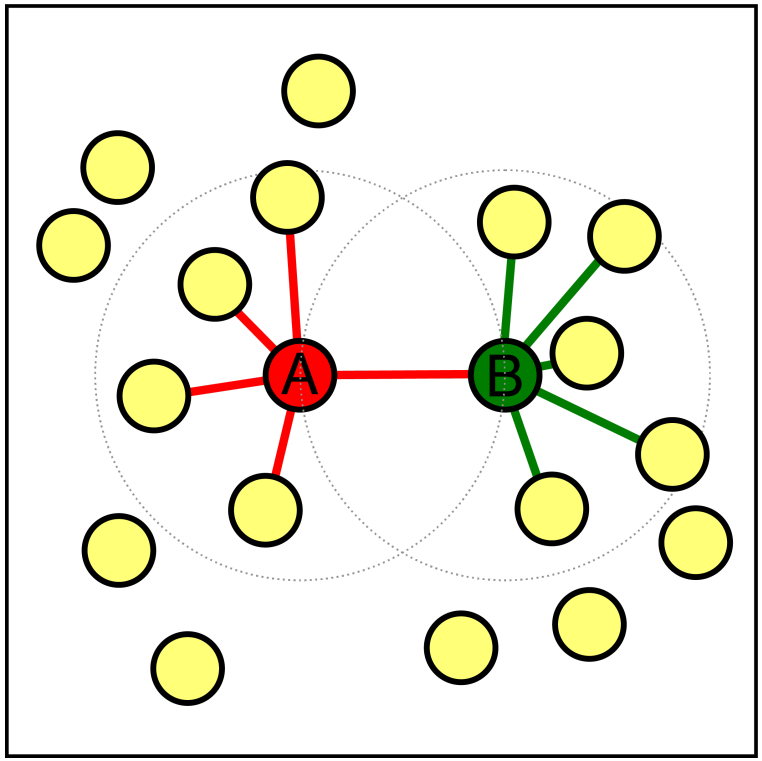
- cells as nodes
- their similarity as edges

In a NN graph two nodes (cells), say A and B, are connected by an edge if:

- the distance between them (in e.g. principal component space) is amongst the **k** smallest distances (here k = 5) from A to other cells, (**K**NN)

or

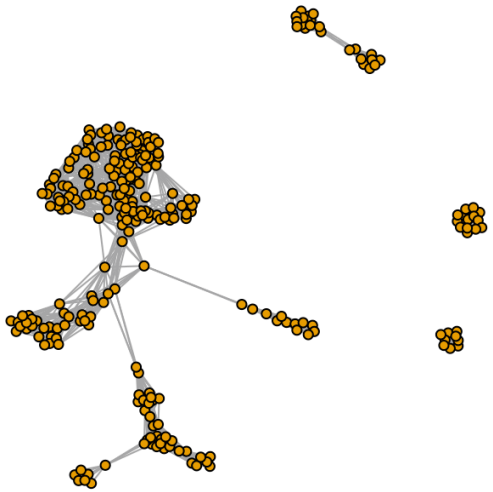
- In a **shared**-NN graph (**S**NN) two cells are connected by an edge if any of their nearest neighbors are shared (n.b. in Seurat this is different)



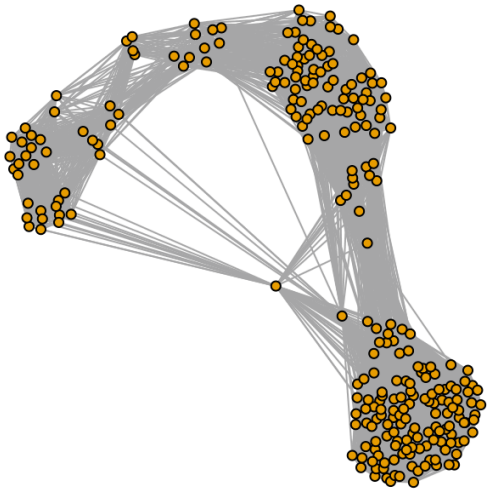
Once edges have been defined, they can be weighted. By default the weights are calculated using the 'rank' method which relates to the highest ranking of their shared neighbours.

Making a graph

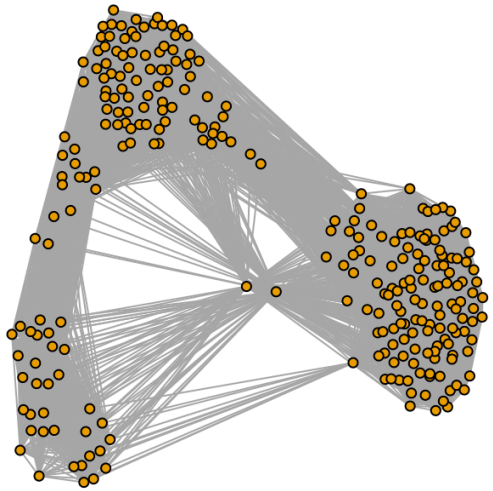
Example with different numbers of neighbours (k):



5-NN



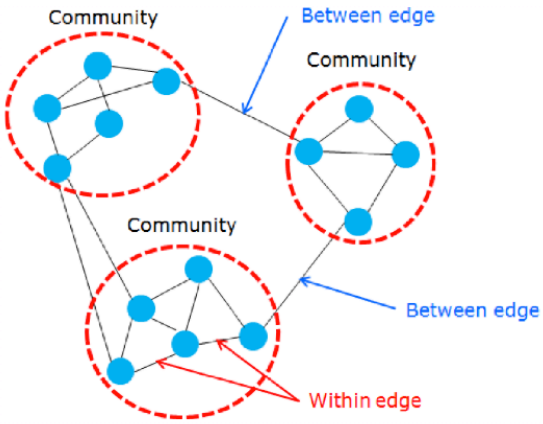
15-NN



25-NN

Identifying communities/clusters

- What makes a community?
  - A community is a cohesive subgroup within a network has following characteristics
    - **Mutual ties**: Most of the members are tied to one another within a community.
    - **Compactness**: A small number of steps are required to reach a group members within a community.
    - **High density of ties**: High density of ties with in a community.
    - **Separation**: High frequency of ties with in a community members when compared to non-members.



community detection algorithms

Here we will address two community detection algorithms: **louvain** and **leiden**.

Modularity

These methods rely on the ‘modularity’ metric to determine a good clustering.

For a given partition of cells into clusters, modularity measures how separated clusters are from each other. This is based on the difference between the observed and expected (i.e. random) weight of edges within and between clusters. For the whole graph, the closer to 1 the better.

Identifying communities/clusters - Louvain

Nodes are also first assigned their own community.

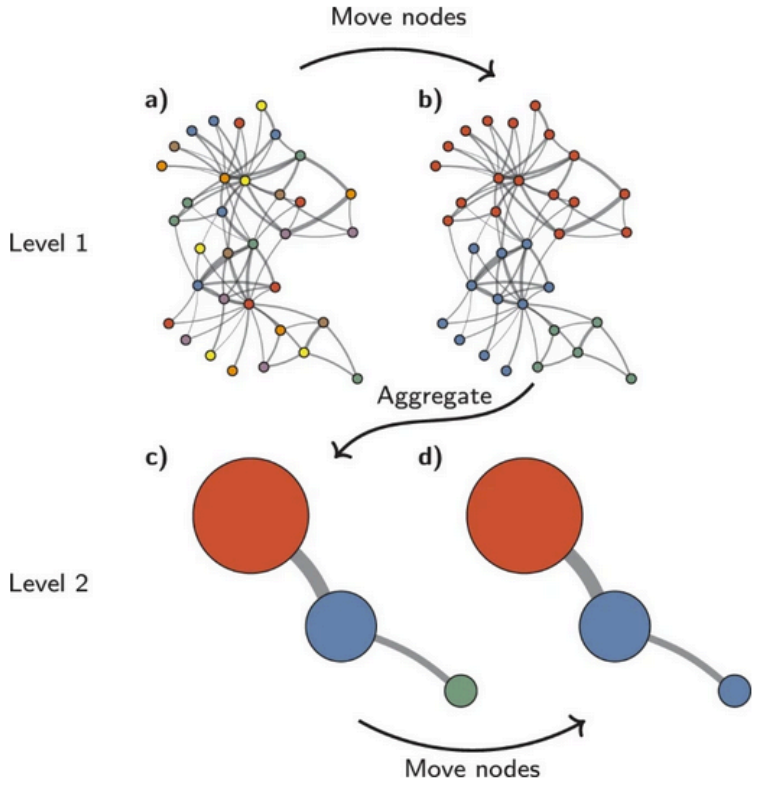
Two-step iterations:

- nodes are re-assigned one at a time to the community for which they increase modularity the most,
- a new, ‘aggregate’ network is built where nodes are the communities formed in the previous step.

This is repeated until modularity stops increasing.

(Blondel et al, Fast unfolding of communities in large networks)

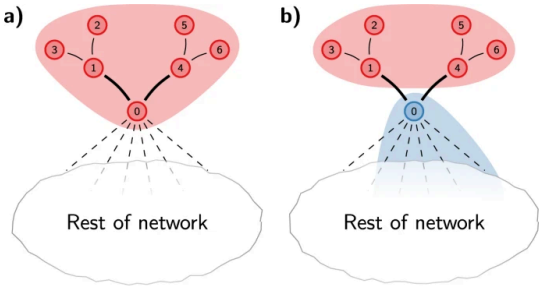
(Traag et al, From Louvain to Leiden: guaranteeing well-connected communities)



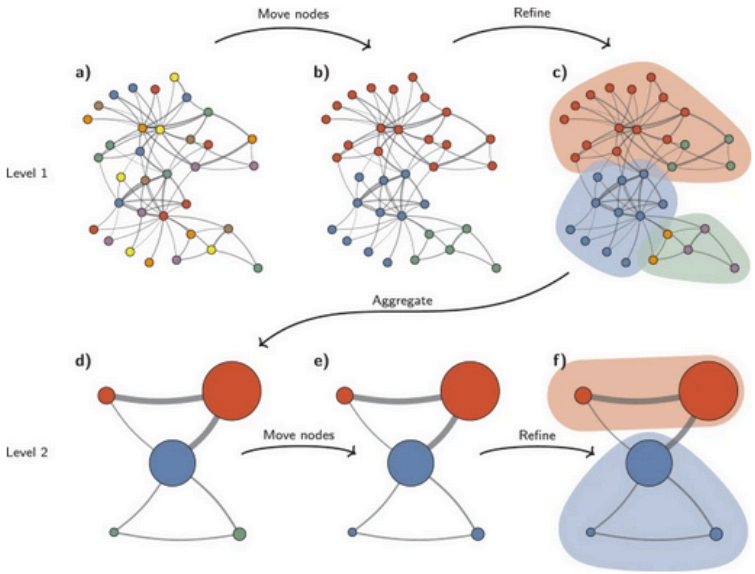
Identifying communities/clusters - Leiden

There is an issue with the Louvain method - some communities may become disconnected.

The Leiden method improves on the Louvain method by guaranteeing that at each iteration clusters are connected and well-separated. The partitioning is refined (step2) before the aggregate network is made.



Disconnected community. Consider the partition shown in (**a**). When node 0 is moved to a different community, the red community becomes internally disconnected, as shown in (**b**). However, nodes 1–6 are still locally optimally assigned, and therefore these nodes will stay in the red community.

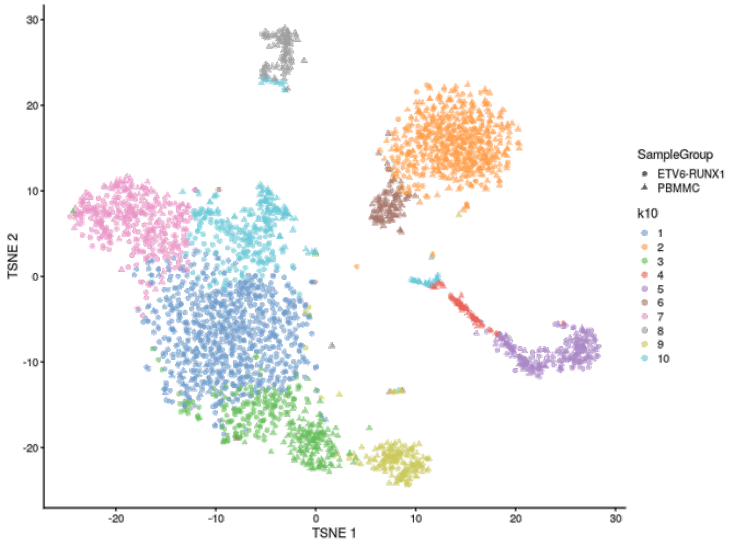


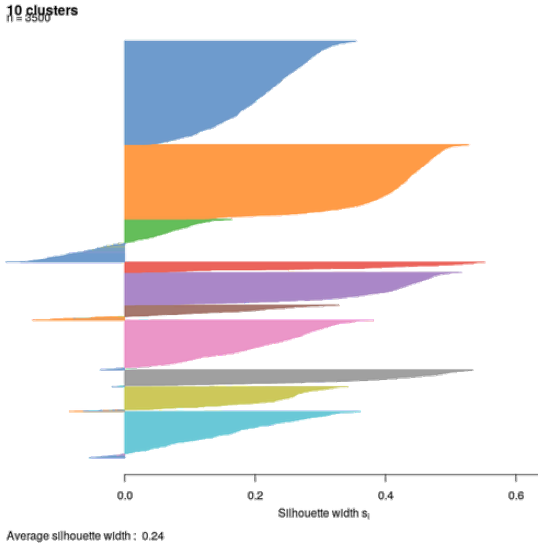
Separatedness - silhouette width

Silhouette width is an alternative to modularity for determining how well clustered the cells are.

```
((mean distance to cells in next closest cluster) - (mean distance to other cells in same cluster))
/ biggest of those means
```

Cells with a large positive width are close to cells in their cluster, while cells with a negative silhouette width are closer to cells of another cluster.





Is there a "correct" clustering?

Clustering, like a microscope, is a tool to explore the data.

We can zoom in and out by changing the resolution of the clustering parameters, and experiment with different clustering algorithms to obtain alternative perspectives on the data.

Asking for an unqualified "best" clustering is akin to asking for the best magnification on a microscope.

A more relevant question is "how well do the clusters approximate the cell types or states of interest?". Do you want:

- resolution of the major cell types?
- Resolution of subtypes?
- Resolution of different states (e.g., metabolic activity, stress) within those subtypes?

Explore the data, use your biological knowledge!